DBSCAN

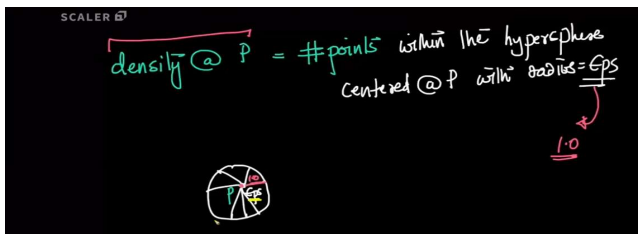
16 February 2024 02:40

- DBSCAN refers to **Density-based spatial clustering of application with noise** .

DBSCAN → Core Point
→ Border Point
→ Outlies / Noise

Run DBSCAN → Hyperparameters → min-no-of pts (3)
→ epsilon (distance) (1 unit)

① Core Point

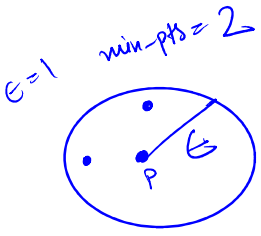


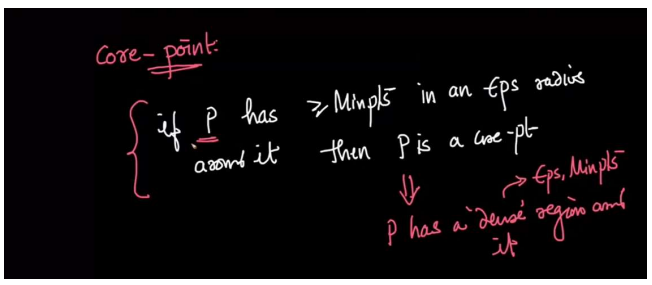
density (P) = 5

min = 2

point ≥ min-pts (2)

Border Point → ① If P is not a core
② & it lies in the neighbourhood of Q
s.t Q is a core point

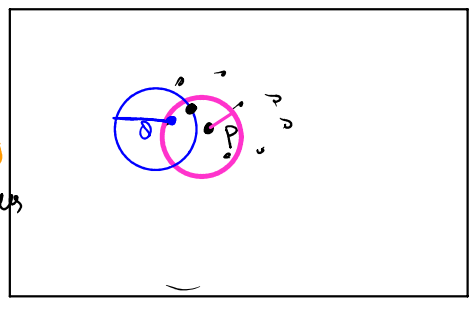




Border Point ϵ = 1
min-pts = 3

P → (4) > 3 ⇒ P is core
Q → (2) ≯ 3 ⇒ Q is border

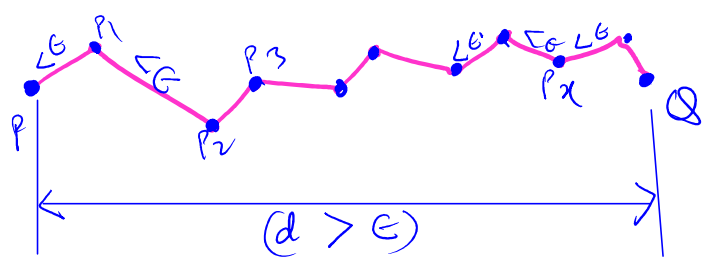
(Q → ✗ core but in vicinity of core)



Outlier → ✗ core
✗ Borders
= outlier

Density Edge : vertex P Q d (edge)
d < Eps (P and Q → core points)

Density connected Points = P and Q (d > Eps but there exists density edge connecting P & Q)



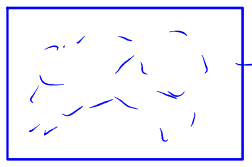
P & Q are core pts

DBSCAN (Algorithm) → Step 1
(ϵ = 1, min pts = 3)

- For each point *xi* that belongs to the dataset *D*, label it as either **core point, border point, or noise point.**

→ core

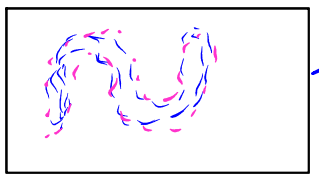
min pts = ...



→ Label each pt as → Core, → Border, → Noise

Step 2

- **Remove all the noise points from the dataset**



Core

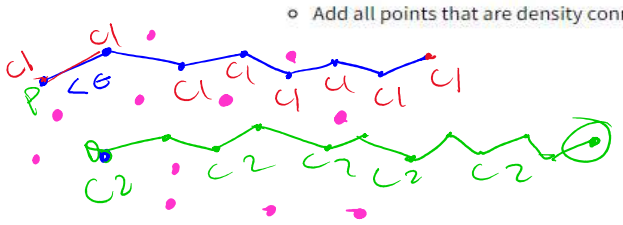
Border

Step 3

Step 3

This step is where things get really interesting. Let's see the 3^{rd} step.

- For each core point P that is not yet assigned to any clustered:
  - create a new cluster with point P
  - Add all points that are density connected to point P, to the P's cluster



Step 4 → Assign Border points to nearest Cluster

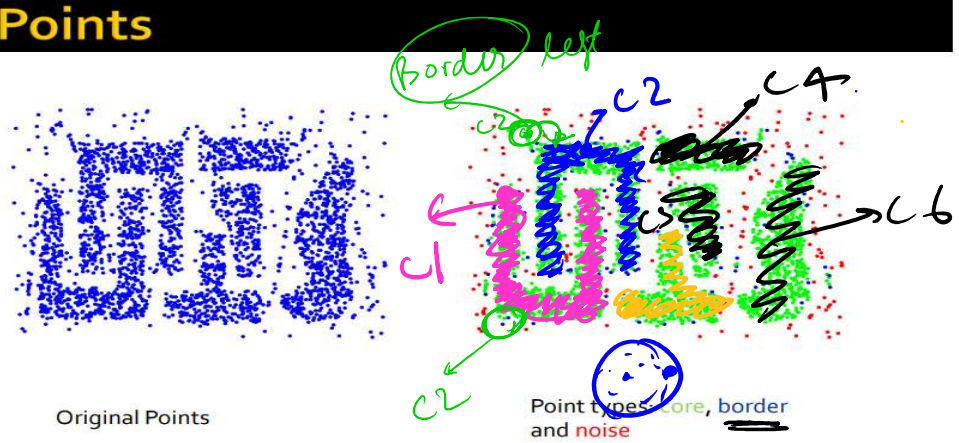
Break till 10:27 pm

Adjusting Hyperparams → Industry → 0.01 (distance) = Telescopic Search

→ min_no_pts ≥ (1D, dimensionality ≥ d+1)

↳ Libraries = (2d) > 20

DBSCAN: Core, Border and Noise Points



Original Points

Point types: core, border and noise

Eps = 10, MinPts = 4

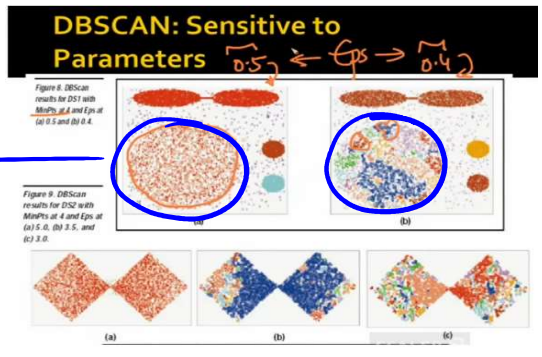
Advantages OF DBSCAN

1. **Resistant to noise**
2. **Handle outliers**
3. **Non linear data as well**
4. **Handle different shapes & sizes**

DISADVANTAGES

1. Sensitive to parameters
2. Distance Based Algo : Not very good with high dimensions

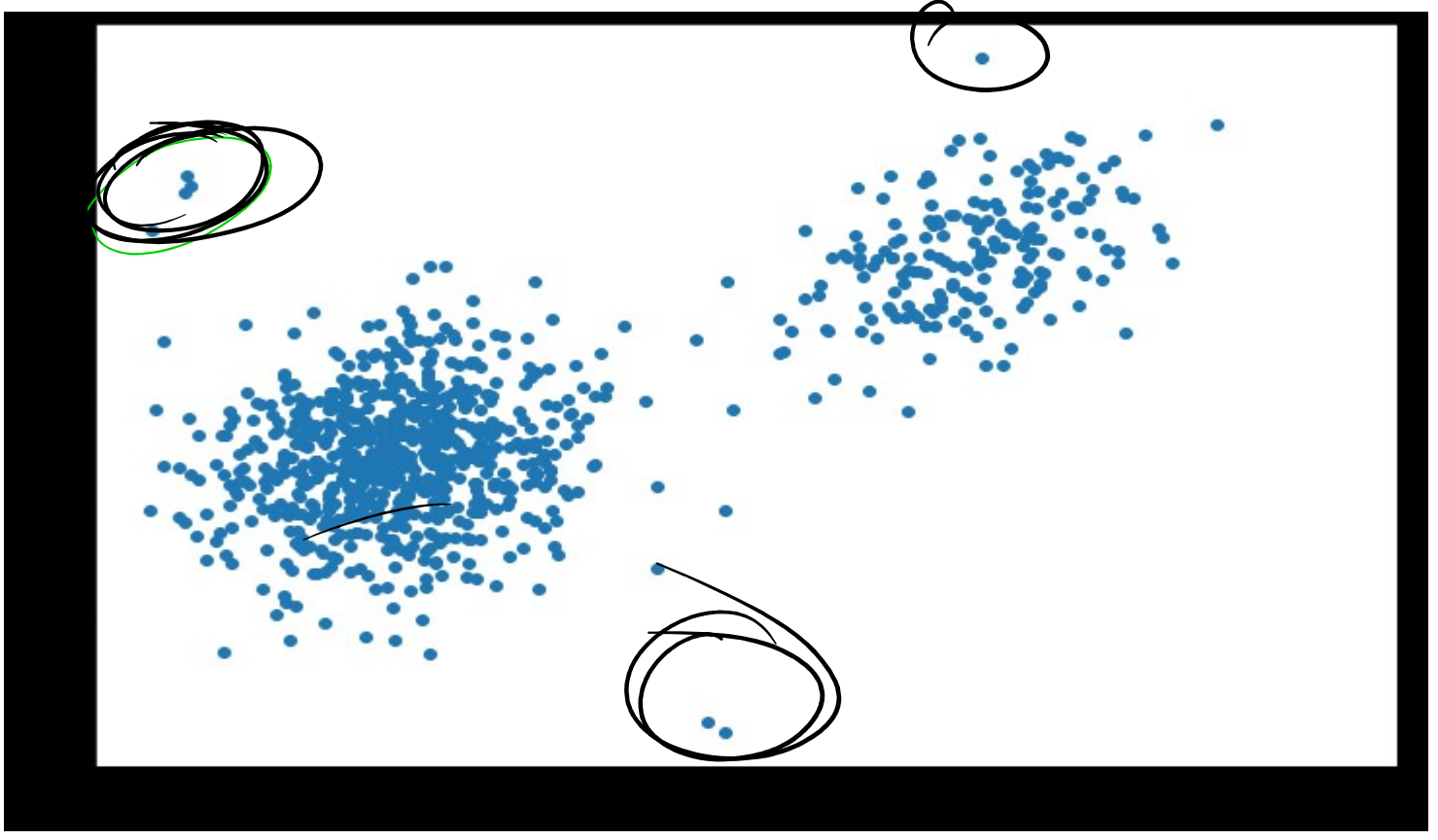
DBSCAN: Sensitive to Parameters



one big clust

Anomaly → outlier, Novelty

Anomaly Outliers



Lokesh → Categorical Data

→ Preprocess → one hot encoding

→ Target encoding

Show them ⟷

Distance Based Metric

① Meaning of Distance vanishes

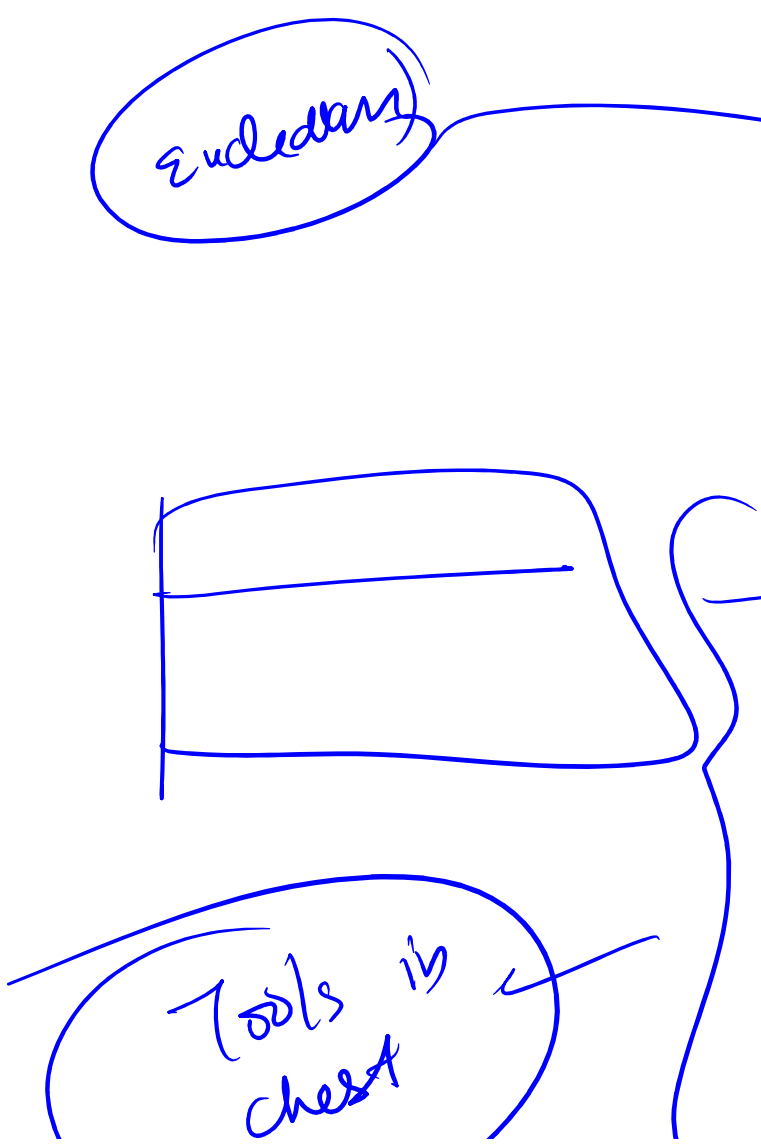
② Custom Distance Metrices

Eucledian →

Full DF Gowers

Hamming Distance Full DF

* (Categorical Data)



→ one hot encod → model.fit

→ Target encoding → y

→ embedding → model.fit

→ Human Data → model.fit

Tools in chest

chest

→ Many Dists → mode-pt

df

d < ∈ → Dists edg

DBSLAN → Distance + Density Based

E – M

posterior probability

Prior

Dirichlet Distrb